Управление в социально-экономических системах
Control in Social and Economic Systems

Научная статья
УДК 004.6
doi: 10.14529/ctcr220107

МЕТОД НАХОЖДЕНИЯ СВЯЗАННЫХ ПОКАЗАТЕЛЕЙ НА ОСНОВЕ АНАЛИЗА НОРМАТИВНО-ПРАВОВЫХ АКТОВ МЕТОДАМИ NLP

***Валерьян Ринатович Аббазов*[1], *Владимир Александрович Балуев*[2], *Андрей Витальевич Мельников*[3], *Михаил Александрович Русанов*[4]**

[1, 2, 3, 4] *Югорский научно-исследовательский институт информационных технологий, г. Ханты-Мансийск, Россия*
[1] *abbazov.v.r2000@gmail.com*
[2] *baluevva@uriit.ru*
[3] *melnikovav@uriit.ru*
[4] *m_rusanov@ugrasu.ru*

Аннотация. Современные методы прогнозирования временных рядов позволяют получить весьма точные и качественные прогнозы при наличии ретроспективных данных. Однако результаты работы этих методов определяются объемом и качеством обучающей выборки. Когда временной ряд отсутствует, имеет малое количество точек или вовсе не достоверен, методы прогнозирования временных рядов неэффективны. В таком случае принято использовать подходы для нахождения иных показателей, так или иначе коррелирующих с искомым, далее называемых косвенными показателями. В рамках работы над прогнозированием социально-экономических показателей возникла необходимость в формировании перечня косвенных показателей, однако имеющиеся решения для данной задачи не обеспечивают требуемой достоверности. В большинстве случаев в работах используют данные социальных сетей, форумов и других источников, которые не могут считаться объективными, так как являются выражением субъективной точки зрения и могут быть подвержены умышленным фальсификациям и искажениям. Такие риски неприемлемы при разработке системы, создаваемой для принятия управленческих решений на уровне государства. **Цель исследования:** разработка методов поиска косвенных показателей, основывающихся на объективных источниках информации. Данные методы позволяют сформировать перечень косвенных показателей, не привлекая экспертов и исключая риски некорректности первичных данных. **Материалы и методы.** Исследования проводились на основе нормативно-правовых актов Российской Федерации и ее субъектов. Данный источник был выбран по причине того, что нормативные документы являются объективными и основополагающими документами государства. Они не являются представлением субъективной точки зрения автора или группы лиц. Для эксперимента была собрана часть нормативной базы с 2016 по 2021 год, относящаяся к категориям: сельское хозяйство, медицина, социальная сфера и другие. **Результаты.** Определен метод нахождения косвенных показателей, разработаны и апробированы различные алгоритмы ранжирования косвенных показателей, сформированы косвенные показатели для нескольких социально-экономических показателей. Процесс выявления косвенных показателей построен на применении методов Data Mining и NLP к базе данных нормативно-правовых актов Российской Федерации. **Заключение.** Полученное решение позволило сформировать список N-грамм, связанных с искомым показателем. На данном этапе интерпретация N-граммы в показатель производится с помощью эксперта, однако для этого не требуется иметь компетенций в предметной области показателя.

Ключевые слова: социально-экономические показатели, N-грамма, показатель деятельности ВДЛ, data mining, NLP

Для цитирования: Метод нахождения связанных показателей на основе анализа нормативно-правовых актов методами NLP / В.Р. Аббазов, В.А. Балуев, А.В. Мельников, М.А. Русанов // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». 2022. Т. 22, № 1. С. 88–96. doi: 10.14529/ctcr220107.

Original article
doi: 10.14529/ctcr220107

METHOD OF FINDING RELATED INDICATORS BASED ON ANALYSIS OF REGULATORY LEGAL ACTS BY NLP METHODS

Valer'yan R. Abbazov[1], Vladimir A. Baluev[2], Andrey V. Melnikov[3], Mikhail A. Rusanov[4]
[1, 2, 3, 4] Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia
[1] abbazov.v.r2000@gmail.com
[2] baluevva@uriit.ru
[3] melnikovav@uriit.ru
[4] m_rusanov@ugrasu.ru

Abstract. Modern methods of forecasting time series allow us to obtain very accurate and high-quality forecasts in the presence of retrospective data. However, the results of these methods are determined by the volume and quality of the training sample. When a time series is missing, has a small number of points, or is not reliable at all, time series forecasting methods are ineffective. In this case, it is customary to use approaches to find other indicators that somehow correlate with the desired one, hereinafter referred to as indirect indicators. As part of the work on forecasting socio-economic indicators, it became necessary to form a list of indirect indicators, however, the available solutions for this task do not provide the required reliability. In most cases, these works use data from social networks, forums and other data sources that cannot be considered objective. Since they are an expression of a subjective point of view and may be subject to deliberate falsifications and distortions. Such risks are unacceptable when developing a system created for making managerial decisions at the state level. **Aim.** Development of methods for searching for indirect indicators based on objective sources of information. These methods make it possible to form a list of indirect indicators without involving experts and eliminating the risks of inaccuracy of primary data. **Materials and methods.** The research was conducted on the basis of regulatory legal acts of the Russian Federation and its subjects. This source was chosen because regulatory documents are objective and fundamental documents of the state. They are not a representation of the subjective point of view of the author or a group of persons. For the experiment, a part of the regulatory framework from 2016 to 2021 was collected, related to the categories: agriculture, medicine, social sphere and others. **Results.** The method of finding indirect indicators is defined, various algorithms for ranking indirect indicators are developed and tested, indirect indicators for several socio-economic indicators are formed. The process of identifying indirect indicators is based on the application of Data Mining and NLP methods to the database of regulatory legal acts of the Russian Federation. **Conclusion.** The resulting solution allowed us to form a list of N-grams associated with the desired indicator. At this stage, the interpretation of the N-gram into an indicator is carried out with the help of an expert, however, this does not require having competencies in the subject area of the indicator.

Keywords: socio-economic indicators, N-gram, VDL activity indicator, data mining, NLP

For citation: Abbazov V.R., Baluev V.A., Melnikov A.V., Rusanov M.A. Method of Finding Related Indicators Based on Analysis of Regulatory Legal Acts by NLP Methods. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics.* 2022;22(1):88–96. (In Russ.) doi: 10.14529/ctcr220107.

Введение

Для решения задачи расчета и прогнозирования показателей эффективности высших должностных лиц необходимо было решить подзадачу сбора первичных данных для расчета показателя и его дальнейшего прогнозирования. В методике расчета данных показателей приведены первичные и их источники, но лишь для некоторых показателей. Первичные данные для оставшихся показателей было необходимо искать в публичных источниках, статистических сборниках и различных отчетах. Для оценки применимости найденных первичных данных было выделено три критерия:

- **объем** – количество точек временного ряда. Чем больше объем, тем лучше модели прогнозирования смогут определить закономерности временного ряда;
- **детализация** – временной интервал между близлежащими точками временного ряда. Так как прогноз показателей деятельности ВДЛ требовался с детализацией по кварталам, то первичные данные с интервалом более года неприменимы (методы интерполяции могли упустить важные закономерности, происходящие в рамках одного года);

● **достоверность** – определяет объективность первичных данных. Достоверными первичными данными являются те, что являются физически исчисляемыми (рождаемость, количество построенных школ, доход населения) и предоставляются специализированными службами, такими как Росстат. В качестве примера недостоверного источника информации можно привести первичные данные, полученные в результате социального опроса, проведенного СМИ. Даже если опрос был проведен корректно, а его результаты могут являться статистически значимыми, то необходимы исследование и проверка корректности каждого такого прогноза, что многократно увеличивает трудозатраты.

В случае если первичные данные соответствовали всем трем критериям, то они были использованы для дальнейших расчетов и прогнозирования. Ожидаемо, что далеко не все первичные данные соответствовали объявленным критериям, более того, для части показателей найти пригодные первичные данные не удалось вовсе.

Для решения задач расчета и прогнозирования показателей, не имеющих первичных данных, было принято решение о разработке метода поиска косвенных показателей, где **косвенный показатель** – это показатель, позволяющий дать оценку динамики изменения для искомого показателя любым другим способом, отличным от описанного в методике расчета [1].

1. Обзор литературы

Анализ публикаций показал, что для задач отбора косвенных показателей, как правило, прибегают к экспертному подходу или рассматривают случай, когда искомый показатель известен и необходимо проверить гипотезу связи искомого показателя с косвенным или набором косвенных показателей [2–7]. В противном случае прибегают к построению графов знаний [8–10], которые позволяют выделить связанные с искомым показателем факторы.

Стоит выделить ряд статей, использующих машинное обучение для интеллектуального анализа текстов для выявления связи расчетного показателя с набором косвенных показателей. Например, в статье [2], в которой описан подход, использующий интеллектуальный анализ текста на основе китайских финансовых новостей в Интернете, для предсказания тенденции цены акций на основе метода опорных векторов. Было обработано более 2 млн новостей в период 2008–2015 годов. С использованием корпуса новостей формируются словарь стоп-слов и точный словарь тональности. На основе описанного корпуса предлагается оригинальная модель прогнозирования с использованием SVM.

Стоит отметить достаточно большое количество статей, посвященных графам знаний. Так, в статье [8] описывается подход для построения тематических графов знаний о мировых событиях на основе газетных статей и показано, что сущности, извлеченные из таких графов, улучшают прогнозы промышленного производства США, Германии и Японии. Для проверки модели использовался корпус из более миллиарда новостных статей за период с 2015 по 2021 год.

Описанные в литературе подходы формирования косвенных показателей можно разделить на три класса: методы на основе экспертной оценки, методы на основе использования графа знаний и методы, использующие инструменты Data Mining.

Экспертный подход обладает рядом недостатков. Во-первых, показатели ВДЛ относятся к слишком разрозненным предметным областям (экономика, здравоохранение, образование, строительство и т. д.), что требует привлечения одного, а лучше нескольких экспертов для анализа каждой предметной области. Во-вторых, экспертный подход является субъективным, при котором невозможно сформулировать строгие критерии отбора косвенных показателей. Не редки случаи, когда эксперты в одной предметной области могут быть не согласны с решениями друг друга. В-третьих, эксперт не всегда способен выявить неявные зависимости. Связь расчетного с косвенным показателем не обязательно может быть интерпретируема.

Основанный на графах знаний подход к построению косвенных показателей также обладает рядом недостатков. Например, в [8, 9] построения графа знаний на основе использования технологий Data Mining описывают процесс обработки новостных заголовков, анализа социальных сетей либо использования глобальных графов знаний, таких как GDELT [11]. Ввиду специфики решаемой задачи основываться на данных из социальных сетей или новостных статей не представляется возможным, так как данные источники не объективны и могут включать в себя умышленные искажения информации.

Помимо этого, большинство подходов, основанных на графах знаний, связывает то, что в качестве вершин графа выступают именованные сущности, полученные в результате NER (Named Entity Recognition) [12–14]. Это не применимо для решаемой нами задачи, так как названия показателей деятельности ВДЛ именованными сущностями не являются. Возможно применить метод построения графа знаний, основывающегося на словах или словосочетаниях, но данный инструментарий не гарантирует получения приемлемого результата в нашем случае, а также сложен и трудозатратен.

В рамках данной работы предлагается использовать методы, которые используют инструменты Data Mining, требующие определения подходящих первоисточников текстовых данных, а также разработки методов агрегации и интерпретации полученных результатов.

2. Модуль извлечения N-грамм, полученных на основе анализа базы данных НПА

В качестве обучающей выборки было решено использовать корпус данных нормативно-правовых актов (НПА) Российской Федерации. Источником получения НПА стал «Официальный интернет-портал правовой информации» [15], являющийся федеральным информационным ресурсом. На данном ресурсе содержатся все правовые акты федерального уровня, уровня субъектов Федерации и муниципальных образований. Данный ресурс насчитывает порядка двух миллионов документов, относящихся к Российской Федерации.

Для проведения эксперимента были использованы только наиболее актуальные нормативно правовые акты. Размер датасета составил более 180 тысяч документов различных тематик, таких как медицина, сельское хозяйство, строительство и прочие.

Для того чтобы сформировать из каждого текста датасета набор N-грамм, был разработан следующий алгоритм.

1. Текст проходит этап предобработки с помощью регулярных выражений, с целью удаления всех символов, кроме русскоязычных букв и знаков пунктуации.
2. Текст разделяется на предложения.
3. Для каждого предложения проводится синтаксический разбор.
4. На основе синтаксического разбора формируются N-граммы из последовательно связанных слов длиной от одного до трех слов.
5. Все слова в N-грамме лемматизируются.
6. Полученная N-грамма заносится в базу данных, сохраняя ссылку на исходный документ.

Разбивка текста на предложения, а также синтаксический разбор предложения осуществлялся с помощью библиотеки Natasha [16], а для лемматизации была использована библиотека pymorphy2 [17].

3. Модуль ранжирования N-грамм

Для того чтобы определить косвенные показатели, необходимо сформировать некое структурированное представление всех полученных ранее 130 миллионов N-грамм. Данное представление должно быть устроено таким образом, чтобы N-граммы с самыми высокими оценками были наиболее семантически близки к названиям косвенных показателей.

В качестве критерия ранжирования была выбрана частотность взаимного упоминания N-грамм. Будем считать, что две N-граммы взаимно упоминаются, если они встречаются в одном тексте. Таким образом, в качестве входного параметра процедура ранжирования получает название целевого показателя, который также преобразуется в N-грамму, а на выходе выдает список всех N-грамм, встречающихся в одном тексте с целевой.

Опишем этапы построения алгоритма ранжирования списка N-грамм, включая его апробацию.

В первую очередь рассмотрим самый простой подход – ранжирование по количеству взаимных упоминаний. При использовании данного подхода N-граммами с самой высокой оценкой будут N-граммы, которые больше всего встречались в одном документе с целевой N-граммой. Преимущество данного подхода в полноте выдачи, так как ни одна N-грамма не будет исключена из итогового списка. Однако это же будет и слабой стороной данного алгоритма, поскольку приводит к слишком высокому зашумлению результатов. При использовании данного алгоритма на вершине выдачи будут появляться такие N-граммы, как «Российская Федерация», «постановле-

ние», «год», «проект» и т. п. Таким образом, наивысшая оценка будет у N-грамм, выполняющих служебную роль в предложении, а не содержательную. Следовательно, необходим алгоритм, позволяющий избавиться от слишком часто употребимых N-грамм, которые присущи практически каждому документу из корпуса НПА и не определяют семантику для расчетного показателя.

В следующей версии алгоритма ранжирования была доработана формула расчета оценки таким образом, что слишком частотные N-граммы, которые встречались во всем корпусе, получали уменьшающий коэффициент, а редкие N-граммы – наоборот, увеличивающий. Данный коэффициент рассчитывался методом аналогично методу TF-IDF [18]. Формула расчета оценки N-граммы выглядела следующим образом:

$$\text{Score} = \left(1 - \frac{\text{Total Mentioning}}{\text{Max Total Mentioning}}\right) \cdot \text{Cross Mentioning}, \quad (1)$$

где Score – оценка ранжирования для N-граммы; Total Mentioning – количество упоминаний N-граммы в корпусе; Max Total Mentioning – количество упоминаний одной N-граммы в корпусе; Cross Mentioning – количество упоминаний N-граммы с целевой N-граммой в одних документах.

При использовании данного подхода слишком частотные N-граммы, аналогично «стоп-словам», получат околонулевые значения. Однако при использовании данного подхода количество неинформативных N-грамм сократилось незначительно. Причем наивысшую оценку имели односоставные N-граммы. Причиной этого стал тот фактор, что, даже имея невысокий множитель, высокочастотная N-грамма встречалась в каждом документе. Поэтому при перемножении коэффициента на количество всех упоминаний, а их может быть больше, чем количество текстов с целевой N-граммой, значения оказывались выше других.

Также проводилась апробация формулы (1) с заменой Max Total Mentioning на Max Cross Mentioning, то есть на максимальную частоту среди документов, содержащих целевую N-грамму, а не всего корпуса. В этом случае качественного улучшения ранжированной выдачи не наблюдалось.

Проанализировав результаты вышеописанных апробаций, было выявлено, что одной лишь оценки, учитывающей частотность N-граммы, недостаточно. Необходимы дополнительные критерии фильтрации, например, такие как добавление порогового значения количества упоминаний N-граммы в корпусе и в документах в целевом множестве.

С учетом всего вышесказанного был предложен следующий алгоритм ранжирования. Множеством всех N-грамм считаются все N-граммы, полученные из корпуса нормативно-правовых актов (рис. 1). Целевым множеством являются N-граммы, встречающиеся с N-граммой искомого показателя в одном документе. Причем N-граммы из целевого множества можно условно поделить на 3 типа.

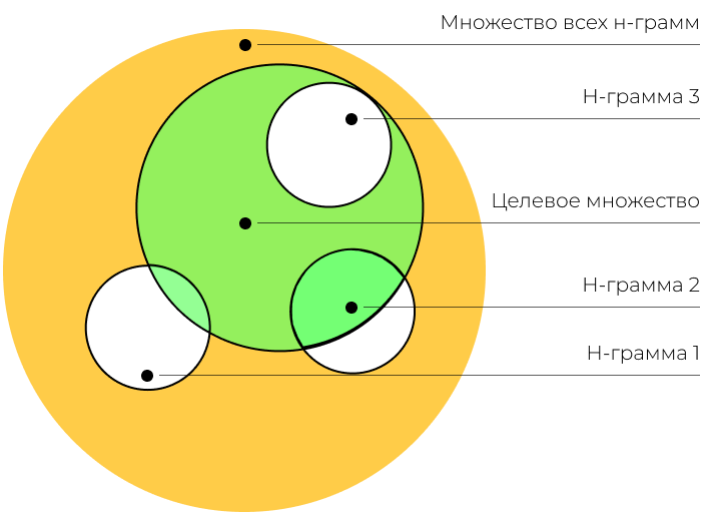

Рис. 1. Множество всех N-грамм
Fig. 1. Set of all N-grams

1. N-грамма почти не встречается в целевом множестве или не встречается вовсе.

2. N-грамма встречается в целевом множестве и в остальной части множества N-грамм.

3. N-грамма встречается только в целевом множестве.

Наиболее значимы при построении расчетного показателя являются N-граммы 2-го и 3-го типов. N-граммы 3-го типа и N-грамма расчетного показателя имеют между собой родительскую связь, а следовательно, это лучшие кандидаты для косвенного показателя, так как вероятнее всего расчетный показатель формируется на основе дочерних. N-граммы 2-го типа интересны в том случае, когда пересечение данных множеств значительно и можно допустить наличие семантической связи между N-граммами.

В предлагаемом алгоритме при ранжировании N-грамм 3-го типа могут встретиться N-граммы, которые имеют низкое количество упоминаний, при этом не являясь содержатель-

ными. Для отсечения подобных случаев необходимо ввести пороговое значение количества упоминаний.

В итоге мы получаем следующий алгоритм ранжирования: для каждой N-граммы, входящей в целевое множество, определяется оценка, равная отношению упоминаний данной N-граммы в целевом множестве к количеству упоминаний данной N-граммы во всем корпусе, при этом исключаются все N-граммы, количество упоминаний которых ниже заданного порогового значения.

Так как пороговое значение существенно зависит от расчетного показателя, то его значение должен определять эксперт опытным путем.

4. Эксперимент

На основе разработанной общей методики с применением созданных программных решений, описанных выше, был проведен эксперимент по прогнозированию показателя деятельности ВДЛ «Численность населения субъекта Российской Федерации».

На основе названия показателя деятельности производим выбор N-граммы: «население, численность».

Производим поиск среди N-грамм собранного датасета НПА и отбираем документы, в которых встречается искомая N-грамма. После этого формируем список всех N-грамм, встречающихся в этих документах. На текущем корпусе НПА для данного показателя деятельности ВДЛ извлечено 3 075 584 N-грамм. Большая их часть малоинформативна, поэтому следующим этапом проводим ранжирование полученного результата.

Эксперт, используя разработанный программный продукт, имеет возможность вручную фильтровать список полученных N-грамм по критерию оценки и количеству упоминаний и выбрать семантически близкие кандидаты для косвенных показателей из списка N-грамм. Пример интерфейса программного продукта представлен на рис. 2.

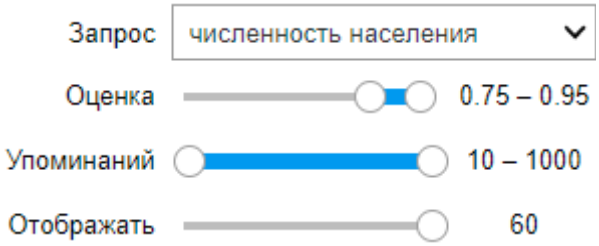

Рис. 2. Пример интерфейса программного продукта
Fig. 2. Example of a product program

Пример полученных N-грамм представлен в таблице.

Пример полученных N-грамм
An example of the obtained N-grams

N-грамма	Количество упоминаний	Оценка
Миграционный, убыль	96	0,989691
Подразделение, расположить, территория	91	0,989130
Посёлок, расположить, территория	90	0,989011
Возраст, дифференциация, норматив	81	0,987805
Возраст, дифференциация	81	0,987805
Обслуживающий, фельдшерский	72	0,986301
Медицинский, организация, подушевой	71	0,986111

Для текущего показателя выделены следующие косвенные показатели:

1) ожидаемая продолжительность жизни при рождении, число лет;
2) число родившихся;
3) число умерших;
4) миграционный прирост, убыль;
5) численность врачей;
6) численность лиц, которым оказана помощь амбулаторно и при выездах.

Выводы

Полученное решение позволило сформировать список N-грамм, связанных с искомым показателем. На данном этапе интерпретация N-граммы в показатель производится с помо-

щью эксперта, однако для этого не требуется иметь компетенции в предметной области показателя.

Стоит отметить, что авторам не известен способ валидации подобных решений. В качестве аргумента в пользу работоспособности данного метода можно привести тот факт, что для показателя «Численность населения» были найдены все указанные в методике расчета показателей ВДЛ переменные: число родившихся, число умерших, показатель миграции. Также в результатах апробации на других показателях ВДЛ были получены логичные, но не очевидные на первый взгляд показатели, например, для показателя «Продолжительность жизни при рождении» был найден косвенный показатель «Численность среднего медицинского персонала» и «Уровень бедности».

Также косвенной валидацией метода поиска связанных показателей на основе анализа нормативно-правовых актов методами NLP может служить объединение прогнозов косвенных показателей и сравнение объединенного прогноза с временным рядом показателя по метрике SMAPE. Для показателей «Население субъекта РФ» и «Ожидаемая продолжительность жизни при рождении» результаты расчета оценки по метрике SMAPE получались 2,3 и 6,6 % соответственно.

Вместе с тем следует отметить важную роль эксперта при интерпретации получаемых N-грамм и выборе кандидатов на косвенные показатели. Снизить субъективизм оценки эксперта возможно путем создания рекомендательной системы, базирующейся на основе метода автоматического определения параметров ранжирования N-грамм.

Список литературы

1. Об утверждении методик расчета показателей для оценки эффективности деятельности высших должностных лиц (руководителей высших исполнительных органов государственной власти) субъектов Российской Федерации и деятельности органов исполнительной власти субъектов Российской Федерации: постановление Правительства Рос. Федерации от 03 апреля 2021 г. № 542. URL: https://docs.cntd.ru/document/560760968 (дата обращения: 21.12.2021).

2. Yancong Xie, Hongxun Jiang. Stock market forecasting based on text mining technology: A support vector machine method. 2019. URL: https://arxiv.org/abs/1909.12789 (дата обращения: 21.12.2021).

3. F. Swen Kuh, Grace S. Chiu, Anton H. Westveld. Modeling National Latent Socioeconomic Health and Examination of Policy Effects via Causal Inference. 2019. URL: https://arxiv.org/abs/1911.00512 (дата обращения: 21.12.2021).

4. Isao Yagi, Yuji Masuda, Takanobu Mizuta. Analysis of the Impact of High-Frequency Trading on Artificial Market Liquidity. 2020. URL: https://arxiv.org/abs/2010.13038 (дата обращения: 21.12.2021).

5. Qi-Qiao He, Patrick Cheong-Iao Pang, Yain-Whar Si. Multi-source Transfer Learning with Ensemble for Financial Time Series Forecasting. 2021. URL: https://arxiv.org/abs/2103.15593 (дата обращения: 21.12.2021).

6. Dilusha Weeraddana, Nguyen Lu Dang Khoa, Lachlan O Neil, Weihong Wang, Chen Cai. Energy consumption forecasting using a stacked nonparametric Bayesian approach. 2020. URL: https://arxiv.org/abs/2011.05519 (дата обращения: 21.12.2021).

7. Rajapaksha D., Bergmeir C., Hyndman R.J. LoMEF: A Framework to Produce Local Explanations for Global Model Time Series Forecasts. 2021. URL: https://arxiv.org/pdf/2111.07001.pdf (дата обращения: 21.12.2021).

8. Sonja Tilly, Giacomo Livan. Macroeconomic forecasting with statistically validated knowledge graphs. 2021. URL: https://arxiv.org/abs/2104.10457 (дата обращения: 21.12.2021).

9. Jie Huang, Kevin Chen-Chuan Chang, Jinjun Xiong, Wen-mei Hwu. Open relation modeling: Learning to define relations between entities. 2021. URL: https://arxiv.org/abs/2108.09241 (дата обращения: 21.12.2021).

10. Madhav Nimishakavi, Uday Singh Saini, Partha Talukdar. Relation schema induction using tensor factorization with side information. 2016. URL: https://arxiv.org/abs/1605.04227 (дата обращения: 21.12.2021).

11. Yihong Yuan. Modeling Inter-country Connection from Geotagged News Reports: A Time-Series Analysis. 2017. URL: https://doi.org/10.1007/978-3-319-61845-6_19 (дата обращения: 21.12.2021).

12. Badgujar A., Chen S., Wang A., Yu K., Intrevado P., Brizan D.G. Quantum Criticism: A Tagged News Corpus Analysed for Sentiment and Named Entities. 2020. URL: https://arxiv.org/abs/2006.05267 (дата обращения: 21.12.2021).

13. Tosin P. Adewumi, Foteini Liwicki, Marcus Liwicki. Word2Vec: Optimal Hyper-Parameters and Their Impact on NLP Downstream Tasks. 2020. URL: https://arxiv.org/abs/2003.11645 (дата обращения: 21.12.2021).

14. Shuxiao Chen, Koby Crammer, Hangfeng He, Dan Roth, Weijie J. Su. Weighted Training for Cross-Task Learning. 2021. URL: https://arxiv.org/abs/2105.14095 (дата обращения: 21.12.2021).

15. Официальный интернет-портал правовой информации. URL: http://pravo.gov.ru/ (дата обращения: 21.12.2021).

16. Veselov D., Kukushkin A., Zamaraev A.N., Yarantsev D., Tihonov S. Solves basic Russian NLP tasks, API for lower level Natasha projects. 2021. URL: https://github.com/natasha/natasha/ (дата обращения: 21.12.2021).

17. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. 2015. URL: https://link.springer.com/chapter/10.1007%2F978-3-319-26123-2_31 (дата обращения: 21.12.2021).

18. Juan Ramos. Using TF-IDF to Determine Word Relevance in Document Queries. 2003. URL: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf (дата обращения: 21.12.2021).

References

1. *Ob utverzhdenii metodik rascheta pokazateley dlya otsenki effektivnosti deyatel'nosti vysshikh dolzhnostnykh lits (rukovoditeley vysshikh ispolnitel'nykh organov gosudarstvennoy vlasti) sub"ektov Rossiyskoy Federatsii i deyatel'nosti organov ispolnitel'noy vlasti sub"ektov Rossiyskoy Federatsii* [On the approval of methods for calculating performance indicators of the supreme executive bodies of state power of the constituent entities of the Russian Federation and the activities of executive bodies of the constituent entities of the Russian Federation]. Available at: https://docs.cntd.ru/document/560760968 (accessed 21 December 2021).

2. Yancong Xie, Hongxun Jiang. *Stock market forecasting based on text mining technology: A support vector machine method*. 2019. Available at: https://arxiv.org/abs/1909.12789 (accessed 21 December 2021).

3. F. Swen Kuh, Grace S. Chiu, Anton H. Westveld. *Modeling national latent socioeconomic health and examination of policy effects via causal inference*. 2019. Available at: https://arxiv.org/abs/1911.00512 (accessed 21 December 2021).

4. Isao Yagi, Yuji Masuda, Takanobu Mizuta. *Analysis of the Impact of High-Frequency Trading on Artificial Market Liquidity*. 2020. Available at: https://arxiv.org/abs/2010.13038 (accessed 21 December 2021).

5. Qi-Qiao He, Patrick Cheong-Iao Pang, Yain-Whar Si. *Multi-source Transfer Learning with Ensemble for Financial Time Series Forecasting*. 2021. Available at: https://arxiv.org/abs/2103.15593 (accessed 21 December 2021).

6. Dilusha Weeraddana, Nguyen Lu Dang Khoa, Lachlan O Neil, Weihong Wang, Chen Cai. *Energy consumption forecasting using a stacked nonparametric Bayesian approach*. 2020. Available at: https://arxiv.org/abs/2011.05519 (accessed 21 December 2021).

7. Rajapaksha D., Bergmeir C., Hyndman R.J. *LoMEF: A Framework to Produce Local Explanations for Global Model Time Series Forecasts*. 2021. Available at: https://arxiv.org/pdf/2111.07001.pdf (accessed 21 December 2021).

8. Sonja Tilly, Giacomo Livan. *Macroeconomic forecasting with statistically validated knowledge graphs*. 2021. Available at: https://arxiv.org/abs/2104.10457 (accessed 21 December 2021).

9. Jie Huang, Kevin Chen-Chuan Chang, Jinjun Xiong, Wen-mei Hwu. *Open relation modeling: Learning to define relations between entities*. 2021. Available at: https://arxiv.org/abs/2108.09241 (accessed 21 December 2021).

10. Madhav Nimishakavi, Uday Singh Saini, Partha Talukdar. *Relation schema induction using tensor factorization with side information*. 2016. Available at: https://arxiv.org/abs/1605.04227 (accessed 21 December 2021).

11. Yihong Yuan. *Modeling Inter-country Connection from Geotagged News Reports: A Time-Series Analysis*. 2017. Available at: https://doi.org/10.1007/978-3-319-61845-6_19 (accessed 21 December 2021).

12. Badgujar A., Chen S., Wang A., Yu K., Intrevado P., Brizan D.G. *Quantum Criticism: A Tagged News Corpus Analysed for Sentiment and Named Entities*. 2020. Available at: https://arxiv.org/abs/2006.05267 (accessed 21 December 2021).

13. Tosin P. Adewumi, Foteini Liwicki, Marcus Liwicki. *Word2Vec: Optimal Hyper-Parameters and Their Impact on NLP Downstream Tasks*. 2020. Available at: https://arxiv.org/abs/2003.11645 (accessed 21 December 2021).

14. Shuxiao Chen, Koby Crammer, Hangfeng He, Dan Roth, Weijie J. Su. *Weighted Training for Cross-Task Learning*. 2021. Available at: https://arxiv.org/abs/2105.14095 (accessed 21 December 2021).

15. *Ofitsial'nyy internet-portal pravovoy informatsii* [Official Internet portal of legal information] Available at: http://pravo.gov.ru/ (accessed 21 December 2021).

16. Veselov D., Kukushkin A., Zamaraev A.N., Yarantsev D., Tihonov S. *Solves basic Russian NLP tasks, API for lower level Natasha projects*. 2021. Available at: https://github.com/natasha/natasha/ (accessed 21 December 2021).

17. Korobov M. *Morphological Analyzer and Generator for Russian and Ukrainian Languages*. 2015. Available at: https://link.springer.com/chapter/10.1007%2F978-3-319-26123-2_31 (accessed 21 December 2021).

18. Juan Ramos. *Using TF-IDF to Determine Word Relevance in Document Queries*. 2003. Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf (accessed 21 December 2021).

Информация об авторах

Аббазов Валерьян Ринатович, программист информационно-аналитического отдела, Югорский научно-исследовательский институт информационных технологий, г. Ханты-Мансийск, Россия; abbazov.v.r2000@gmail.com.

Балуев Владимир Александрович, руководитель центра информационно-аналитических систем, Югорский научно-исследовательский институт информационных технологий, г. Ханты-Мансийск, Россия; baluevva@uriit.ru.

Мельников Андрей Витальевич, д-р техн. наук, проф., директор, Югорский научно-исследовательский институт информационных технологий, г. Ханты-Мансийск, Россия; melnikovav@uriit.ru.

Русанов Михаил Александрович, старший преподаватель института цифровой экономики, Югорский научно-исследовательский институт информационных технологий, г. Ханты-Мансийск, Россия; m_rusanov@ugrasu.ru.

Information about the authors

Valer'yan R. Abbazov, programmer of Information Analysis Department, Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia; abbazov.v.r2000@gmail.com.

Vladimir A. Baluev, head of the Center for Information and Analytical Systems, Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia; baluevva@uriit.ru.

Andrey V. Melnikov, Dr. Sci. (Eng.), Prof., director, Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia; melnikovav@uriit.ru.

Mikhail A. Rusanov, senior lecturer of the Institute of Digital Economy, Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia; m_rusanov@ugrasu.ru.

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации.

Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article.

The authors declare no conflicts of interests.

Статья поступила в редакцию 23.12.2021; одобрена после рецензирования 12.01.2022; принята к публикации 18.01.2022.

The article was submitted 23.12.2021; approved after reviewing 12.01.2022; accepted for publication 18.01.2022.